DOCUMENT RESUME

ED 205 577 TM 810 466

AUTHOR Yeh, Jennie P.; Conklin, Jon
TITLE Using the Rasch Model to Examine Item Bias.
INSTITUTION California Univ., Los Angeles. Center for the Study of Evaluation.
SPONS AGENCY National Inst. of Education (ED), Washington, D.C.
REPORT NO CSE-R-151
PUB DATE Nov 80
NOTE 35p.

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Computer Programs; *Goodness of Fit; Grade 4; Intermediate Grades; *Latent Trait Theory; Mathematical Models; Measurement Techniques; *Socioeconomic Status; *Test Bias; *Test Items
IDENTIFIERS *Rasch Model

ABSTRACT

To test for item bias, it must be determined whether an item fits the model. Two approaches to defining bias within the framework of the Rasch model are examined. One compares within-group fit mean squares and the other utilizes a between-group fit statistic. Results from both approaches overlap somewhat, but are distinct in many different but complimentary, and both may be useful to the analyst interested in both aspects of the bias issue. These two indices are by no means the only indices to be used within the Rasch model framework. Apparently, a promising alternative approach might focus on person-fit rather than item-fit. (Author/GK)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED205577

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

# USING THE RASCH MODEL TO EXAMINE ITEM BIAS

Jennie P. Yeh
Jon Conklin

CSE Report No. 151

November, 1980

TM 810 466

Center for the Study of Evaluation
Graduate School of Education, UCLA
Los Angeles, California 90024

The research reported herein was supported in whole or in part by a grant to the Center for the Study of Evaluation from the National Institute of Education, U. S. Department of Education. However, the opinions and findings expressed here do not necessarily reflect the position or policy of NIE and no official NIE endorsement should be inferred.

## Table of Contents


| | Page |
|---|---|
| Introduction | 1 |
| The Rasch Model | 2 |
| The Study | 4 |
| Results | 9 |
| Entire Sample | 9 |
| Within Socioeconomic Groups | 13 |
| Between Socioeconomic Groups | 16 |
| Conclusions | 18 |
| References | 25 |

# INTRODUCTION

Increasingly, latent trait models have shown promise for application in the areas of test construction, item pooling, test equating, and tailored testing. One of these models, the Rasch model (Rasch, 1960, 1966), has enjoyed much popularity because of several advantages it has over other techniques. In addition to its simplicity, both in theory and in application, it is the only latent trait model which enables sample-free estimation of person as well as item parameters. Recent discussions (Hambleton, et al., 1978) have suggested that the Rasch model may hold promise for examining test bias. There are several clear advantages to such an approach over classical test theory for defining and identifiying bias. The Rasch model can be used to identify biased items not just biased tests. It requires no assumptions about the comparability of ability distributions of different groups or about within-group reliabilities. In addition, its conclusions are not dependent upon the characteristics of the sample of persons taking the test since estimates of item characteristics are sample invariant. Unfortunately, few studies have actually used the Rasch model in examining bias, and thus precise definitions of "bias" are unclear. Items are generally identified as biased if they exhibit a lack of fit to the model characterizing the test as a whole. The Rasch model itself makes the strong assumptions that items must represent a single, unidimensional, underlying trait, and that item discriminations are equal. An item's lack of fit to the model essentially indicates that one of these assumptions is being violated. The item may represent different traits for different persons, or it may discriminate between persons in a manner unlike other items on the test. Either of these interpretations could be construed to conform to a general definition of bias.

According to this approach, then, to test for item bias we need only determine whether an item fits the model. Several tests of item fit have been suggested (Gustaffson, 1979). We will focus on a simple technique proposed by Wright and Mead (1978) which essentially measures the average squared deviations between obtained and predicted item characteristic curves. The statistic has a mean of one and can be tested for statistical significance. It is included as a part of the BICAL program (Wright, 1977), for Rasch model calibrations. Similar to analysis of variance decompositions, this index of total item-fit is made up of between-group and within-group components. Definitions of item bias may be based on either one of these fit statistics. In his examination of the issue of bias, Durovic (1975) suggests that significant differences in within-group-fit mean squares for a given item is evidence that the item is biased with respect to those groups. Essentially, this amounts to testing for group by fit interactions. An alternative definition of bias could be based on the between-group-fit mean square. Items with significant between-group-fit mean squares may be interpreted as testing different traits in different groups, or as differing from the remaining items in terms of how they discriminate between groups. Though the conventional approach in this test of item-fit for identifying groups is to form them on the basis of ability (total score), we are more concerned with socioeconomic and racial groups. By applying the same tests of fit to such groupings we can identify items that are biased in terms of socioeconomic status or race.

## THE RASCH MODEL

The Rasch Model assumes that items are dichotomously scored, the test is not speeded, and that the odds for success can be defined as a

function of the ratio of person ability ($\beta_v$) to item difficulty ($\delta_i$)

$$O_{vi} = \frac{\beta_v}{\delta_i} \qquad [1]$$

where $O_{vi}$ is the odds for person v to succeed on item i, $\beta_v$ is the ability of person v, and $\delta_i$ is the difficulty of item i. If the probability of obtaining a correct response is defined as the odds for success divided by one plus the odds, we obtain the following:

$$P_{vi} = \frac{O_{vi}}{1+O_{vi}} = \frac{\beta_v/\delta_i}{1 + \beta_v/\delta_i} \qquad [2]$$

We now have an expression for the Rasch probability of a correct response in terms of only two parameters, person ability and item difficulty. To make the model simpler by putting it into an additive form we define $\beta_v$ as the log ability of person v and $\delta_i$ as the log difficulty of item i. The probability of a correct response can then be expressed as:

$$P_{vi} = \frac{e^{(\beta_v-\delta_i)}}{1 + e^{(\beta_v-\delta_i)}} \qquad [3]$$

When person v has more of the latent ability than item i requires, then $\beta_v$ exceeds $\delta_i$ and the probability of success is greater than 0.5. If the item is too difficult for person v, then $\delta_i$ exceeds $\beta_v$ and the probability is less than 0.5.

In contrast to other latent trait models, the Rasch model specifies only one item parameter, difficulty. Other models also use an exponential function of person and item parameters to define the probability of a successful response, but specify additional item parameters of discrimination and tendency to provoke guessing. The Rasch model essentially sets the guessing parameter to zero and treats the discrimination parameter as

if it were constant for all items. Naturally, these are strong assumptions. However, the measurement logic they defend can be supported (Wright, 1977).[1] As a direct result of these assumptions, "The Rasch model is the only latent trait model for a dichotomous response that is consistent with 'number right' scoring" (Wright, 1977). Furthermore, it is the only method for both obtaining estimates of item parameters free of the ability distribution of the person sample and estimates of person parameters free of the difficulty distribution of the item sample.

## THE STUDY

In this study, the Rasch model was used to examine the scores obtained on a fourth-grade, 31-item arithmetic test administered as part of a large scale evaluation of compensatory educational programs. A total of 1007 fourth grade students were sampled from California elementary schools to represent a cross-section of socioeconomic and program strata. The content covers the skills of basic computation in all four operations, word problems, and fractions. Items are scored dichotomously, the test has no restrictive time limit for completion and an internal consistency reliability of .88.

Of the several available computer programs for applying the Rasch model, the BICAL program developed by Wright and Mead (1978) was used because it includes a number of features that are not included in other programs. In addition, it incorporates a test of item-fit which produces between-group, within-group, and total fit mean squares. BICAL defines

---

[1] These assumptions may be viewed not so much as restrictive assumptions which must be met prior to applying the model, but rather as ideals on which the model is based. They act primarily to define item-fit.

groups on the basis of total score, forming up to six of these score groups (based on user specifications). The basic criterion is that groups have approximately equal sizes. The fit statistics consist, basically, of residuals from the model in terms of item difficulty both for observations within a score group and for the separate score group means. The between-group fit statistic tests whether observed item characteristic curves for separate groups have a common shape and slope. As stated in the BICAL program manual:

> If estimates of difficulty are in fact free of the distribution of ability in the calibrating sample, then estimates based on different subgroups will be statistically equivalent to those based on the total sample. This can be tested most severely by dividing the sample into subgroups based on score level and each item in each score group with those predicted for that subgroup from the total sample estimate (Wright & Mead, 1978, p. 12).

Within-group fit is essentially an extension of this logic to a comparison of each person-item interaction to the expected value of an item's characteristics based on that person's group as a whole. This decompositon of item-fit is analogous to the partitioning of sums of squares in the analysis of variance.

Interpreting an item's lack of fit depends, to a certain extent, on whether the lack of fit occurs between or within groups. An item with a significantly large between-group fit mean square is not discriminating among ability groups in the same manner as the remaining items on the test. That is, groups of lower ability may be more successful and groups of high ability less successful (or vice versa) than expected given their performance on the rest of the test and the resulting model predictions. Of course, this can be taken as evidence that the item is testing different skills or trait dimensions at different ability levels; that is, the item may violate the unidimensionality

assumption of the model. On the other hand, one could argue that the same trait is being measured though item discrimination differs. Both interpretations imply that the item is biased with respect to the groups examined relative to the other items on the test. An item with a significantly large within-group fit mean square, on the other hand, may not necessarily be biased, especially if it does not lack between-group fit. Such a case would be evidence to the effect that though not biased between groups, smaller gradations of ability are not consistently detected by the item. This may indicate that certain characteristics of the item, possibly unrelated to its content or the underlying trait dimension, may be ambiguous or confusing. Such an item may be of abnormal form or length, be too novel, or be poorly constructed.

Mathematically, the fit statistics are calculated in the same manner as conventional mean squares (e.g., in ANOVA). Squared standardized residuals (between obtained values and model predicted values) are summed and divided by the appropriate degrees of freedom. The between-group mean square compares the successes in group g on item i, $S_{gi}$, to their model expectation:

$$S_{gi} \approx \sum_{r \epsilon g} n_r m_i P_{ri} \qquad [4]$$

where $n_r$ is the number of persons obtaining a score of r, and $P_{ri}$ is the estimated probability of success given the ability estimate $b_r$ associated with a score of r and the difficulty estimate $m_i$ associated with item i (Wright & Mead, 1978, p. 8).[2] The $r \epsilon g$ specifies that the terms $n_r$ and

[2] In the actual expressions, a term $m_i$ is included for replications. Here each person interacts with each item once; thus $m_i$ has been set to one.

$P_{ri}$ and the summation are only for observation within-group g. The full between-group mean square can be expressed as:

$$V_{Bi} = \sum_{g} \left[ \frac{(S_{gi} - \sum_{r \epsilon g} n_r m_i P_{ri})^2}{\sum_{r \epsilon g} n_r m_i P_{ri}(1-P_{ri})} \right] \cdot \left[ \frac{L}{(g-1)(L-1)} \right] \qquad [5]$$

This statistic is distributed with an expected value of 1.0 and a variance of $2L/[(g-1)(L-1)]$ where L equals total number of items. Naturally, this can be further expanded by substituting for $P_{ri}$ as in expression [3].

The within-group mean square is obtained by comparing the between-group statistic to the total mean square which is expressed as:

$$V_i = \sum_{v=1}^{n} \left[ \frac{(X_{vi} - P_{vi})^2}{P_{vi}(1-P_{vi})} \right] \cdot \left[ \frac{L}{(n-1)(L-1)} \right] \qquad [6]$$

where N is the total number of persons and $X_{vi}$ is the result of a specific person-item interaction.

Examining, now, the two definitions of item bias presented earlier, it is apparent that the difference between the statistics mentioned there and the statistics just presented above concerns the method of forming groups. In defining bias, fit statistics must be computed based on groups for which the issue of bias is relevant. Such groups might be formed on the basis of race (Durovic, 1975) or socioeconomic status. These groups, of course, overlap in score distributions and thus cannot be directly formed through the BICAL program without major program alterations. Durovic's method of comparing total fit mean squares calculated separately for each group requires only that separate BICAL runs be made for each group. Comparing the total fit statistic obtained in such a manner for each group would be

similar to comparing the within-group squared standardized residuals produced on a single BICAL run[3] in which groups are formed on the basis of an outside criterion as desired. The approach identifies as biased those items which fit the model significantly better in one group than in another. Rather than comparing the deviations of item behavior per group from overall model predictions, each item's behavior within a group is compared to model predictions based on that group alone. An alternative definition of item bias is suggested when we consider that for an item to be unbiased

> [the] item characteristic curves which provide the probabilities of correct responses must be identical across different sub-populations of interest (Hambleton, et al., 1978, p. 94).

This implies that between-group rather than within-group statistics should be compared to model predictions based on all groups combined. Though Durovic's approach does make comparisons between groups, it merely compares the within-group item fits to each group's model predictions. This latter approach actually involves the calculation of a between-group fit statistic which describes how item behavior at the group level differs from an overall model prediction. The statistic involved is actually the same between-group fit mean square presented in expression [5] except that groups are formed on the basis of an outside criterion rather than on the basis of total score. Rather than make the extensive program revision necessary to enable BICAL to form groups and calculate statistics on the basis of an outside criterion, all of the necessary values can be obtained if separate runs are obtained for each group (here, based on socioeconomic status) and one is obtained for all groups combined.[4]

---

[3]The program has not been set up to independently calculate and print such statistics per group.

[4]The value for $P_{ri}$ is based on the combined-groups estimates of $b_4$ and $d_i$, whereas $S_{gi}$ and $n_r$ are based on information provided in each of the separate groups run. The between-group mean square can be then calculated for each item outside the BICAL program itself.

In the following sections we examine, first the item characteristics of the test for the entire sample. Items lacking fit (based on score groups) in the entire sample are identified and examined. Groups are then formed on the basis of socioeconomic status, and separate analyses are performed on each group. Within-group mean squares are then computed, and between-group fit mean squares are calculated. The findings, using both methods of defining item bias, are discussed and contrasted. The BICAL program developed by Wright and Mean (1978) is used for all analyses.

## RESULTS

### Entire Sample

As a first step, the Rasch model was applied to the entire sample of 1007 fourth-graders. The data for each subject consisted of the 31 dichotomously scored (wrong-right) items. Because perfect scores and zero scores provide no item information, the Rasch model excludes persons obtaining such scores from all analyses. In this sample, 17 persons answered all 31 items correctly and one person answered all items incorrectly, thus leaving 989 persons for item calibration.

Table I presents the difficulties and fit statistics estimated for each item. It should be made clear that the difficulty scale is somewhat arbitrary. The difficulty scale reported is expressed in logits, with a mean of zero and with positive values indicating above average difficulty, negative values indicating below average difficulty. We can see that the easiest items are items 1, 2, 3, 8, and 20. These items were answered correctly by 91, 89, 86, 85, and 81 percent of the subjects respectively. Examining the content of those items in Appendix A we see that the first three are simple, straightforward addition problems. Item 8 is a simple

multiplication problem without carrying, and item 20 is a word problem requiring simple addition. Apparently these skills are fairly well mastered by most fourth-graders.

Examining the difficult items, we see that items 17, 31, 30, and 21 were most difficult in that order. They were answered correctly by only 26, 28, 33, and 35 percent of the subjects respectively. Item 17 represents the only "complex" division problem presented in the test. It consists of a multiple digit divisor and requires "long division" (the only other long division problem was answered correctly by only 41% of the subjects). Examination of the common errors failed to reveal any noticeable patterns. Items 30 and 31 both represent the only examples of reexpressing fractions. Errors on both were usually made in a consistent direction: "1/2" was thought to be equal to "2/3," and "8/10" was thought to be equal to "7/9." That is, subjects apparently attended to the size of the difference between the value of the denominator and the value of the numerator. Item 27 represents the only item requiring the subtraction of complex fractions (a whole number with a fraction). The common errors were on responses B and D, both of which are also complex fractions. In conclusion, we can say that for this sample of fourth-graders, long division problems and problems involving fractions are most difficult.

In Table II the items with significant total fit mean squares are presented in order of their fits. Recall that total fit actually consists of two orthogonal components and provides an overall index of how well an item fits the model describing the test as a whole. As previously stated, each fit statistic is distributed with a mean of one and the standard error can be estimated based on the number of items, subjects, and groups. Items with significantly large fit mean squares represent items that do not fit

the model; that is, there is a discrepancy between the obtained item characteristics curves and those predicted knowing the behavior of the test as a whole and the number of persons correctly responding to each item. Such a discrepancy indicates that a number of subjects did not respond as predicted by the model. Examining the table we see that of the 11 poorly fitting items, three are from the most difficult items and three are from the easiest items. Therefore, it seems there is no clear relationship between item difficulty and item fit. Of the 11 poorly fitting items, we see that four represent problems dealing with fractions and five represent simple addition, subtraction, or multiplication problems. These items with poor total fits would be deleted from the test since they fail to behave in a manner consistent with the test as a whole. By forming ability groups of approximate equal size (the program ranks out 6 ability groups) total fit can be broken into orthogonal components. With respect to the resulting within-groups and between-group fits, there are two features worth noting in our results. First, several items exhibit very large between-group fit mean squares. Some of these, because they exhibit good within-group fit, do not have large total fit mean squares. We see from Table I that 13 out of the 31 items have fit statistics greater than 3 standard errors from the mean. Though not explicitly presented here, an examination of the average responses by score groups provides insight into the nature of these poor between-group fits. In general, the lower score groups performed worse than expected on multiplication items and word problems, but better than expected on certain subtraction and division items and on fractions. A possible interpretation of these patterns could be as follows: Subtraction and division problems may be uniformly difficult for all children regardless of ability and thus may not easily

discriminate on the basis of ability. In addition, students of lower ability may be provided with extra practice and training on such difficult concepts. The poorer performance of low ability students on word problems probably reflects a lower reading ability and thus poorer comprehension of item meanings. Their much better than expected performance on fractions may reflect a tendency for teachers to monitor such students more closely and provide them with more feedback than they would with students of higher ability. Alternatively, it may be that because the actual computations required on the fraction problems are quite simple, cumulative knowledge may not limit the performances of lower ability students.

A second point concerns those items with large within-group but small between-group fit squares. Such a pattern implies that whereas different ability groups are conforming on the average to the model, persons within groups are not. This may indicate that certain item features are confusing or require great concentration or attention; such features are likely to result in much variation from person to person but have little effect on group means. That is, such items, though not necessarily violating model assumptions, are poor from the standpoint of introducing unwanted within-group variability. As can be seen from Table I, most items with large within-group mean squares also lack between-group fits, thus implying that they have violated model assumptions. Items which have significant within-group fits tend also to have significant between-group fits. Items 1, 7, and 8 exhibit such a pattern. Though item 1 is the easiest item, the fact that it is first on the test may have resulted in random errors merely because students are in a rush to get started. Item 7 is made up of multiple tasks and thus requires much concentration in that it is lengthy and requires carrying. Item 8 is the first

nonaddition or substraction item and thus may be confusing to some students. Items with large within-group but small between-group fits may be characterized by much random guessing.

In summary, then, a conventional Rasch model analysis of the entire sample, with item-fit being based on score groupings, shows that in this 31-item test of fourth grade arithmetic skills, a number of items appear to behave poorly. These items are primarily problems dealing with fractions or simple operations. Their formats, or the underlying skills which they call for, may result in guessing. Thus, they fail to discriminate between score groups in the manner in which other test items do. For test review purposes, such results would indicate that these items should be deleted or rewritten before the final draft of the test.

Within Socioeconomic Groups

The BICAL program has no provisions for creating groups on the basis of an outside criterion (e.g., socioeconomic status [SES]) and thus, within SES-group fit, could not be examined using the same approach as the one described above. Instead, separate SES-group files were created, and Rasch model analyses were performed separately for each. The total fit mean-square obtained for items in a specific SES-group would then be equivalent to the within-group mean squares had fit been examined in the conventional manner with SES-groupings. Table III provides the basic fit statistics and difficulty estimates for the items with significantly large total fits within each of the SES groups. After perfect and zero scorers were excluded, the low, middle, and high SES groups were represented by 428, 348, and 213 subjects respectively.

Examining the table we see that the groups differ in the number of non-fitting items and in the orders of item-fits. Only items 1 and 31 appear to be consistently poor in all three groups. It should be mentioned that whereas item 31 is the most difficult item for the middle group, for the low SES group item 17 is the most difficult. Recall earlier statements to the effect that lower scorers did better than expected on fraction problems, and note that SES and test performance are generally highly correlated. Of course direct comparisons of item difficulties across groups cannot accurately be made since they have not been standardized. Slight scale differences may be present.

It is apparent that some of the non-fitting items are group specific. That is, items may fit in certain groups and not in others. This type of pattern has been taken by Durovic (1975) as evidence of item bias--differential within-group fits. Table IV presents non-fitting items according to their differential lack of within-group fit. As stated previously, items 1 and 31 lack fit in each of the groups and according to Durovic's definition are not necessarily biased items. Items 17 and 25 fit in the high SES group but not in the middle and low SES groups. Other items show different patterns of fit and non-fit. As a first step toward interpreting these patterns, we should recognize that schools generally reflect the characteristics of their surrounding neighborhoods. That is, schools tend to be much more homogeneous with respect to SES than with respect to student ability. Thus patterns of differential within-group fits may provide an indication of differential school effects. For instance, the complex division represented in item 17 may be emphasized more in higher socioeconomic schools, and thus may better conform in behavior with the remaining test items for that group. The fact that item 25 is a word

problem dealing with fractions may mean that responses to a large extent depend on the ability to read and understand the item stem itself. Children from high SES home and school backgrounds may have received greater support for reading activities (their parents are generally more educated) and thus be less likely to be confused by the reading content of such an item. Examining items 3 and 4 we may conjecture that in lower SES schools the more basic operations are emphasized and thus such items would have discriminability similar to the remaining items within the low SES group. In the higher SES groups such skills may be dealt with in less detail and repetition, and thus longer or newly formatted items such as numbers 3 and 4 may elicit more confusion and guessing. An especially interesting item is number 19, for it is the worst fitting item in the high SES group but fits well in the lower SES groups. Examination of within-group patterns shows that within the high SES group, lower ability persons do better than expected, and higher ability persons do worse than expected. That is, the item appears to be almost uniformly easy for persons of differing abilities. In the middle and lower SES groups, persons who have lower total scores (ability) tend to do substantially worse on this item than do persons with higher total scores. Thus in these groups the item appears to fit the model.

To be sure, the interpretations made above are not the only viable ones that could be made. However, it is likely that the SES group differences that they do represent are school level phenomena. One possible school level effect that may make a difference is the differential exposure to certain concepts or skills. That is, an item may fit because all students have been exposed to the concepts contained in it; and thus ability is the primary determining factor for success or failure on that

item. On the other hand, an item may not fit if students have been exposed only briefly and thus random guessing is common. An item also may not fit if a skill has been mastered by the majority of the subjects, thus making the item uniformly easy. To the extent that test scores may reflect differences in exposure to concepts, and that such exposure may be SES related, the comparison in this section of within-group fits as suggested by Durovic (1975) may be a legitimate exercise for identifying item bias.

Between Socioeconomic Groups

If the previous interpretation of between-group lack of fit is correct--that non-fitting items discriminate between groups in a manner inconsistent with the rest of the test--then we might surmise that by grouping individuals on the basis of SES, we could run a single Rasch model analysis and use between-group fit as an index of bias. Unfortunately, the BICAL program does not enable one to form groups on the basis of an outside criterion. Of course, we could act as if ability were a proxy for SES and present the earlier findings concerning the entire sample as our examination of bias. On the other hand, though SES is highly correlated with ability, the score distributions of the three SES groups examined here are highly overlapping. Thus, between-ability-group fits may not be consistent with between-SES-group fits. Fortunately, the actual formula used for calculating between-group fit is straightforward (Wright and Mead, 1978), and the necessary values can be obtained if separate analyses have been performed on each SES group as well as on the entire sample.

The between-SES-group fit mean squares are presented in Table V for all 31-items on the test. The statistic has an expected value of one and a standard error of 1.02. Thus all mean squares greater than 3.04 represent non-fitting (or biased) items. Items 31, 23, 4, 28, 19, and 26 are identified by this procedure as biased with respect to socioeconomic status. The fact that they don't fit the model indicates that they may tap different underlying traits in different SES groups or may discriminate between groups in a manner inconsistent with the test as a whole. Some of these items have been identified and discussed before. Specifically, items 31, 4, and 19 have been identified as non-fitting items in the analysis of the entire sample, and with the addition of item 26, have been identified in the examination of within-SES-group fit. It is interesting to note that whereas item 1 does not fit in the entire sample and consistently lacks within-SES-group fit, it appears to fit relatively well between groups. Also interesting is the fact that items 23 and 28 lack between-SES-group fit but appear to fit in all previous analyses. Item 31 is identified as biased using between-group mean squares, whereas it was not specified as biased using the within-group approach. Many of the items identified as biased in the within-group analyses do not appear to be biased when using the between-group definitions.

Examining item contents, we see that three of the six biased items represent problems dealing with fractions. It appears that lower SES students are performing higher than expected on such problems, whereas higher SES students are performing lower than expected. This pattern is also true in the case of item 4 which is a column addition problem. On items 19 and 23, just the opposite is true: high SES students are performing better than expected and low SES students are performing worse.

Thus we may conclude that with respect to these items, emphasizing word problems in arithmetic tests may exaggerate the apparent SES differences, whereas problems dealing with fractions and (to a certain extent) simple operations may minimize such differences in total scores. The evidence that items 31 and 4 have large within-group as well as between-group fit mean squares indicates that they should be deleted from the test regardless of the implications of their bias. Certainly, other items should also be examined even if they are not biased, because of their lack of within-group fit.

## CONCLUSIONS

Use of the Rasch model for examining item bias seems to have great potential. In this study, we have examined two different approaches to defining bias within the framework of the Rasch model. One compares within-group fit mean squares and the other utilizes a between-group fit statistic. Results from both approaches overlap to a certain extent, but are distinct in many respects. The indices of bias they provide are slightly different but complementary, and both may be useful to the analyst interested in both aspects of the bias issue.

It is important to note that any definition of bias which rests on the use of item-fit statistics falls prey to a fundamental problem. Fit is a relative measure. It merely measures deviation of items from the test as a whole. It is true that this is a problem in classical approaches as well as in those using the Rasch model; but the possibility remains that an item lacking fit may actually be a "good" item while the test as a whole is "poor."

Finally, it should be noted that the two indices of bias examined here are by no means the only indices one could use within the Rasch model framework. It appears that a promising alternative approach might focus on person-fit rather than item-fit. A measure of overall person-fit calculated for a specific group of persons would indicate the extent to which items (or groups of items) were behaving as we would expect from model predictions. Gustafsson (1979) has suggested that person-fit measures may indeed be the only way to examine test unidimensionality--lack of unidimensionality that varies across persons may be evidence of test bias. Further work in this area may be promising.

Table I

Item Analysis of Entire Sample
(Fit Statistic Based On Six Ability Groups)

| | | FIT MEAN SQUARE | | | | |
|---|---|---|---|---|---|---|
| | Item Diff | Withn Group | Betwn Group | Total | Disc Indx | Point Biser |
| 1 | -2.31 | 1.36 | 0.53 | 1.36 | 0.97 | 0.26 |
| 2 | -1.96 | 1.04 | 0.91 | 1.03 | 1.12 | 0.36 |
| 3 | -1.63 | 1.22 | 2.55 | 1.22 | 0.95 | 0.34 |
| 4 | -0.91 | 1.43 | 6.76* | 1.47 | 0.77 | 0.32 |
| 5 | -0.64 | 0.97 | 0.66 | 0.96 | 1.03 | 0.46 |
| 6 | -0.86 | 0.93 | 0.81 | 0.93 | 1.04 | 0.44 |
| 7 | 0.02 | 1.13 | 0.72 | 1.13 | 0.94 | 0.47 |
| 8 | -1.51 | 1.47 | 2.07 | 1.47 | 0.86 | 0.30 |
| 9 | -0.22 | 0.88 | 2.85 | 0.89 | 1.15 | 0.53 |
| 10 | -0.20 | 0.94 | 1.51 | 0.95 | 1.05 | 0.50 |
| 11 | 0.75 | 0.81 | 4.52 | 0.84 | 1.28 | 0.59 |
| 12 | 0.61 | 0.87 | 3.70 | 0.89 | 1.23 | 0.56 |
| 13 | -0.10 | 0.76 | 7.18 | 0.80 | 1.34 | 0.60 |
| 14 | 0.01 | 0.81 | 2.72 | 0.83 | 1.20 | 0.55 |
| 15 | 1.11 | 1.01 | 1.43 | 1.01 | 1.05 | 0.50 |
| 16 | 0.83 | 0.83 | 4.64* | 0.86 | 1.28 | 0.58 |
| 17 | 1.96 | 1.26 | 5.56* | 1.29 | 0.73 | 0.36 |
| 18 | -0.24 | 1.02 | 0.82 | 1.01 | 1.08 | 0.50 |
| 19 | -1.00 | 1.19 | 2.99* | 1.20 | 1.04 | 0.44 |
| 20 | -1.22 | 0.79 | 2.92* | 0.80 | 1.16 | 0.47 |
| 21 | -0.42 | 0.77 | 4.72* | 0.80 | 1.26 | 0.56 |
| 22 | 0.54 | 1.00 | 0.55 | 1.00 | 1.07 | 0.51 |
| 23 | 0.86 | 0.86 | 4.17* | 0.88 | 1.25 | 0.57 |
| 24 | -0.46 | 1.21 | 2.18 | 1.22 | 0.83 | 0.40 |
| 25 | 0.29 | 1.30 | 4.56* | 1.32 | 0.73 | 0.40 |
| 26 | 0.32 | 1.04 | 1.31 | 1.04 | 0.90 | 0.45 |
| 27 | 1.42 | 1.11 | 1.07 | 1.11 | 0.89 | 0.44 |
| 28 | 0.75 | 1.02 | 0.95 | 1.02 | 0.94 | 0.49 |
| 29 | 0.85 | 1.03 | 0.64 | 1.03 | 0.90 | 0.47 |
| 30 | 1.54 | 1.06 | 2.35 | 1.06 | 0.95 | 0.45 |
| 31 | 1.86 | 1.54 | 20.38* | 1.65 | 0.42 | 0.23 |
| | | 983 | 6 | 989 | DEG OF FROM | |
| | | 0.05 | 0.58 | 0.04 | STD ERROR | |

*items with between group fits greater than 3SE from the mean.

Table II

Items with Significant Lack of Fit)

N = 989

| Number | Total Fit (.04) | W/in grp. (.05) | Betw. grp. (.53) | Diff. |
|---|---|---|---|---|
| 31 | 1.65 | 1.54 | 20.33 | 1.86 |
| 8 | 1.47 | 1.47 | 2.07 | -1.51 |
| 4 | 1.46 | 1.43 | 6.76 | -0.91 |
| 1 | 1.35 | 1.36 | 0.53 | -2.31 |
| 25 | 1.31 | 1.30 | 4.56 | 0.29 |
| 17 | 1.28 | 1.26 | 5.56 | 1.96 |
| 3 | 1.22 | 1.22 | 2.55 | -1.63 |
| 24 | 1.22 | 1.21 | 2.18 | -0.46 |
| 19 | 1.19 | 1.19 | 2.99 | -1.00 |
| 7 | 1.13 | 1.13 | 0.72 | 0.02 |
| 27 | 1.10 | 1.11 | 1.07 | 1.92 |

Table III:

Items with Significant Lack of Fit Within Separate SES Groups

| LOW SES (N=428) | | | MID SES (N=348) | | | HI SES (N=213) | | |
|---|---|---|---|---|---|---|---|---|
| # | Total Fit | Diff. | # | Total Fit | Diff. | # | Total Fit | Diff. |
| 31 | 1.64 | 1.49 | 4 | 1.70 | -1.06 | 19 | 2.67 | -1.60 |
| 17 | 1.50 | 2.09 | 3 | 1.59 | -1.91 | 8 | 2.23 | -1.22 |
| 25 | 1.37 | 0.26 | 1 | 1.57 | -2.69 | 31 | 1.58 | 2.04 |
| 1 | 1.32 | -2.23 | 31 | 1.50 | 2.23 | 4 | 1.53 | -0.39 |
| 24 | 1.23 | -0.52 | 25 | 1.31 | 0.22 | 24 | 1.46 | -0.59 |
| 8 | 1.16 | -1.71 | 17 | 1.24 | 1.88 | 2 | 1.46 | -1.95 |
| | | | 18 | 1.20 | -0.40 | 7 | 1.41 | 0.06 |
| | | | 27 | 1.17 | 1.63 | 3 | 1.29 | -1.27 |
| Standard Error | .07 | | | .08 | | 26 | 1.24 | 0.61 |
| | | | | | | 1 | 1.20 | -1.95 |

Table IV

Items with Differential Lack of Fit Across SES Groups
(+ means fit; - means lack of fit)

| Item | Low | Mid | High |
|---|---|---|---|
| 1 | - | - | - |
| 31 | - | - | - |
| 17 | - | - | + |
| 25 | - | - | + |
| 4 | + | - | - |
| 3 | + | - | - |
| 8 | - | + | - |
| 24 | - | + | - |
| 18 | + | - | + |
| 27 | + | - | + |
| 19 | + | + | - |
| 2 | + | + | - |
| 7 | + | + | - |
| 26 | + | + | - |

Table V

Item Analysis of Entire Sample
(Fit Statistics Based on Three SES Groups)

| Item | Departure from expected ICC | | | Between group fit mean square |
|---|---|---|---|---|
| | High SES | Middle SES | Low SES | |
| 1 | -.02 | .01 | -.01 | 1.96 |
| 2 | -.01 | .01 | -.01 | .66 |
| 3 | -.03 | .02 | -.00 | 2.30 |
| 4 | -.06 | .03 | .02 | 5.54 |
| 5 | -.03 | .00 | .03 | 1.54 |
| 6 | -.02 | .01 | .02 | .88 |
| 7 | -.00 | .03 | -.01 | 1.13 |
| 8 | -.03 | -.01 | .03 | 2.35 |
| 9 | .03 | .01 | -.01 | 1.00 |
| 10 | .00 | .01 | .01 | .21 |
| 11 | .02 | .01 | -.02 | .78 |
| 12 | .05 | -.03 | .01 | 2.50 |
| 13 | .02 | .02 | -.01 | .82 |
| 14 | .04 | .03 | -.02 | 2.55 |
| 15 | .03 | -.04 | .01 | 2.04 |
| 16 | .02 | -.01 | .00 | .21 |
| 17 | -.01 | .01 | -.02 | 1.12 |
| 18 | .03 | .04 | -.03 | 2.99 |
| 19 | .06 | .00 | -.02 | 3.65 |
| 20 | .01 | .01 | -.01 | .33 |
| 21 | .02 | .02 | -.01 | 1.01 |
| 22 | .04 | .01 | -.02 | 1.53 |
| 23 | .08 | .01 | -.05 | 6.83 |
| 24 | .02 | -.01 | .02 | .99 |
| 25 | -.03 | .03 | .01 | 1.36 |
| 26 | -.05 | .00 | .04 | 3.14 |
| 27 | .00 | -.03 | .02 | 1.60 |
| 28 | -.01 | -.04 | .04 | 3.78 |
| 29 | -.03 | .00 | .01 | .72 |
| 30 | -.01 | -.01 | .00 | .16 |
| 31 | -.04 | -.06 | .05 | 9.17 |

REFERENCES

Durovic, J. J. Test bias: An objective definition for test items. Paper presented at the 1975 Annual Convocation of the Northeastern Educational Research Association, 1975.

Gustafsson, J. E. Testing and obtaining fit of data to the Rasch model. Paper presented at the 1979 Annual Meeting of the American Educational Research Association, San Francisco, 1979.

Hambleton, R. K., et al. Developments in latent trait theory: Models, technical issues, and applications. Review of Educational Research, 1978, 48, 467-510.

Linn, R. L., & Werts, C. E. Considerations of test bias. Journal of Educational Measurement, 1971, 8, 1-4.

Rasch, G. Probabilistic models for some intelligence and attainment tests. Copenhagen, Denmark: Denmarks Paedogogiske Institut, 1960.

Rasch, G. An item analysis which takes individual differences into account. British Journal of Mathematical and Statistical Psychology, 1966, 19(1), 49-57.

Wright, B. D. Sample-free test calibration and person measurement. Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton, NJ: Educational Testing Service, 1968.

Wright, B. D. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977, 14, 97-116.

Wright, B. D., & Mead, R. J. BICAL: Calibrating items and scales with the Rasch model, (Research Memorandum No. 23A). Statistical Laboratory, Department of Education, University of Chicago, 1978.

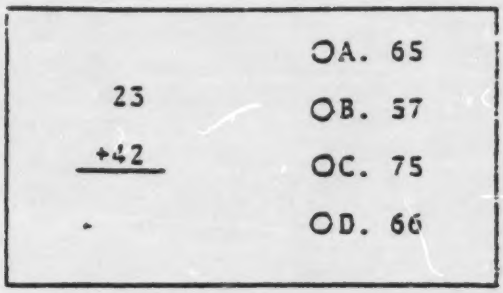
NAME: _____

GRADE: _____

ROOM: _____

SCHOOL: _____

Example A: Do the math problem. Fill in the circle next to the correct answer.




Example B: Read the problem, and figure out the answer. Fill in the circle next to the correct answer.

Mark had 6 marbles. John gave him 3 more marbles. How many marbles does Mark now have?

OA. 3 OB. 10 OC. 9 OD. 6

Do each math problem. Fill in the circle next to the correct answer. Fill in only one circle for each problem.

**Example**

$$\begin{array}{r} 12 \\ +\ 3 \\ \hline \end{array}$$

○A. 12
○B. 9
○C. 15
○D. 6

"C" is the correct answer

| | | | | | |
|---|---|---|---|---|---|
| 1. | $\begin{array}{r} 69 \\ +\ 22 \\ \hline \end{array}$ | ○A. 81<br>○B. 91<br>○C. 87<br>○D. 47 | 5. | 94 - 78 = | ○A. 26<br>○B. 24<br>○C. 16<br>○D. 14 |
| 2. | 66 + 35 = | ○A. 101<br>○B. 91<br>○C. 102<br>○D. 99 | 6. | $\begin{array}{r} 47 \\ -\ 28 \\ \hline \end{array}$ | ○A. 29<br>○B. 25<br>○C. 19<br>○D. 21 |
| 3. | $\begin{array}{r} 3357 \\ +\ 2447 \\ \hline \end{array}$ | ○A. 5794<br>○B. 5804<br>○C. 6814<br>○D. 5704 | 7. | $\begin{array}{r} 6600 \\ -\ 2573 \\ \hline \end{array}$ | ○A. 4127<br>○B. 4173<br>○C. 4027<br>○D. 4137 |
| 4. | $\begin{array}{r} 144 \\ 35 \\ 221 \\ +\ 73 \\ \hline \end{array}$ | ○A. 363<br>○B. 473<br>○C. 474<br>○D. 373 | 8. | $\begin{array}{r} 23 \\ \times\ 3 \\ \hline \end{array}$ | ○A. 66<br>○B. 69<br>○C. 59<br>○D. 56 |

GO ON TO NEXT PAGE

9.

$$\begin{array}{r} 86 \\ \times\ 7 \\ \hline \end{array}$$

OA. 562
OB. 616
OC. 602
OD. 596

10.

$$\begin{array}{r} 193 \\ \times\ 6 \\ \hline \end{array}$$

OA. 1148
OB. 1049
OC. 1158
OD. 648

11.

$$\begin{array}{r} 12 \\ \times\ 34 \\ \hline \end{array}$$

OA. 308
OB. 418
OC. 48
OD. 408

12.

$$\begin{array}{r} 402 \\ \times\ 16 \\ \hline \end{array}$$

OA. 2874
OB. 412
OC. 6432
OD. 6492

13. $8\overline{)32}$

OA. 4 Remainder 2
OB. 5
OC. 4
OD. 3 Remainder 4

14. $4\overline{)30}$

OA. 6
OB. 8
OC. 7 Remainder 2
OD. 7 Remainder 5

15. $8\overline{)336}$

OA. 54
OB. 42
OC. 40 Remainder 6
OD. 32

16. $25\overline{)75}$

OA. 3
OB. 5 Remainder 4
OC. 3 Remainder 4
OD. 4

17. $15\overline{)255}$

OA. 21 Remainder 5
OB. 17
OC. 10 Remainder 5
OD. 16

GO ON TO NEXT PAGE

Read each problem. Fill in the circle next to the correct answer.

18. One thousand ten is:

○A. 1,010 ○B. 10,010 ○C. 1,001 ○D. 100,010

19. Five hundred four is:

○A. 5,040 ○B. 504 ○C. 5,004 ○D. 540

20. Patty read 27 pages of her book before lunch and 8 pages after lunch. How many pages did she read?

○A. 19 ○B. 36 ○C. 35 ○D. 45

21. Sue had 48 marbles. If she gave 22 marbles to Juan, how many did she have left?

○A. 26 ○B. 16 ○C. 70 ○D. 25

22. Tria rode her bicycle in the country at 12 miles an hour for 3 hours. How many miles did she go in that time?

○A. 4 ○B. 36 ○C. 15 ○D. 46

23. Bob had 42 baseball cards. He made 6 piles with his cards and was sure to put the same number of cards in each pile. How many cards were in each pile?

○A. 7 ○B. 52 ○C. 6 ○D. 48

24. This circle is divided into equal parts. What part of the circle is shaded?

○A. $\frac{8}{3}$ ○B. $\frac{3}{8}$ ○C. $\frac{2}{4}$ ○D. $\frac{2}{6}$

25. If a student answers 18 problems correctly out of 20, what proportion of problems did she answer correctly?

○A. $\frac{2}{20}$ ○B. $\frac{18}{20}$ ○C. $\frac{20}{18}$ ○D. $\frac{20}{2}$

GO ON TO NEXT PAGE

Do each fraction problem. Fill in the circle next to the correct answer. Fill in only one circle for each problem.

Example

$\frac{2}{6} + \frac{3}{6} =$

○A. $\frac{5}{12}$

●B. $\frac{5}{6}$

○C. $\frac{5}{8}$

○D. $\frac{6}{12}$

"B" is the correct answer

26. $\frac{10}{12} - \frac{5}{12} =$

○A. $\frac{5}{24}$

○B. $\frac{5}{12}$

○C. $\frac{7}{12}$

○D. 5

27. $8\frac{1}{3} - \frac{1}{3} =$

○A. 8

○B. $7\frac{2}{3}$

○C. 7

○D. $8\frac{1}{6}$

28. $\frac{3}{7} + \frac{2}{7} =$

○A. $\frac{5}{7}$

○B. $\frac{5}{14}$

○C. $\frac{6}{14}$

○D. $\frac{14}{21}$

29. $3\frac{5}{8} + 2\frac{2}{8} =$

○A. $6\frac{7}{8}$

○B. $5\frac{7}{8}$

○C. $5\frac{7}{16}$

○D. $5\frac{16}{40}$

30. $\frac{1}{2} = \frac{?}{\quad}$

○A. $\frac{2}{3}$

○B. $\frac{2}{5}$

○C. $\frac{4}{9}$

○D. $\frac{2}{4}$

31. $\frac{8}{10} = \frac{?}{\quad}$

○A. $\frac{3}{4}$

○B. $\frac{4}{5}$

○C. $\frac{7}{9}$

○D. $\frac{3}{5}$

STOP